Microbiology (2005), **151**, 2125–2132 **DOI** 10.1099/mic.0.28006-0

Mini-Review

Databases and software for the comparison of prokaryotic genomes

Dawn Field,[1] Edward J. Feil[2] and Gareth A. Wilson[1]

[1]Oxford Centre for Ecology and Hydrology, Mansfield Road, Oxford OX1 3SR, UK

[2]Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

Correspondence
Dawn Field
dfield@ceh.ac.uk

The explosion in the number of complete genomes over the past decade has spawned a new and exciting discipline, that of comparative genomics. To exploit the full potential of this approach requires the development of novel algorithms, databases and software which are sophisticated enough to draw meaningful comparisons between complete genome sequences and are widely accessible to the scientific community at large. This article reviews progress towards the development of computational tools and databases for organizing and extracting biological meaning from the comparison of large collections of genomes.

No genome is an island

Genomes can never be considered as isolated datasets or 'islands'. Rather, they must be viewed and interpreted in the context of the large amount of molecular data available in public databases. We now have a vast collection of genomes in our public databases. This collection contains more than 2500 genomes from bacteria (>200), viruses (>1200), plasmids (>600), eukaryotes (>30) and organelles (>500) (Field *et al.*, 2005; Wheeler *et al.*, 2005). While there are numerous resources for the study of genomes we focus here on databases and software for the study of complete prokaryotic genomes.

First-generation tools focused on the study of a single genome, and the majority of genomic resources developed to date identify features in individual genomes expressly for the sake of functional annotation. Now there is an obvious shift towards the creation of tools that allow the viewing and manipulation of data in a comparative genomic context. While many of these 'next-generation' comparative genomics databases and software packages combine information from multiple sources to 'decorate' a single target genome with finer detail, there are also truly comparative genomic tools that leverage the information in multiple genomes simultaneously. Tools that allow the direct comparison of two or more genomes are becoming increasingly common (Darling *et al.*, 2004; Hohl *et al.*, 2002; Kurtz *et al.*, 2004). For example, the Artemis Comparison Tool (Table 1) provides a visualization of BLAST hits between two complete genome sequences, thus allowing rapid examination of the degree of synteny (conservation in gene order), major genomic rearrangements, or the integration of novel genomic islands, phages or other 'foreign' elements.

Analysis of evolutionarily and ecologically richer collections of complete genomes

Access to genomes from closely related species accelerates the functional annotation of novel genomes (Kurtz *et al.*, 2004). Likewise, biologically interesting features often only become apparent in a comparative genomic context. Such features include orphans, gene fusions and pseudogenes. For example, the ability of gene prediction programs to correctly identify pseudogenes depends on the quantity and type of mutations accumulated in these evolutionarily 'dead genes'. The successful annotation and interpretation of the *Mycobacterium leprae* genome depended on the detection of a large number of pseudogenes in the unusually numerous non-coding regions. Many had decayed so far that they could only be identified through comparison with the complete genome sequence of the larger, more metabolically flexible relative *Mycobacterium tuberculosis*, which contains functional homologues (Cole *et al.*, 2001). Comparative genomic studies have also led to a growing appreciation of the mosaic nature of bacterial genomes and the unexpectedly high levels of genetic diversity found among bacterial isolates corresponding to a single named bacterial 'species'.

Despite all the advantages of continually growing collections of genomes, increased taxonomic and ecological richness itself is not enough to solve all the challenges associated with interpreting the contents of these genomes. We simultaneously need to improve the quality and speed of annotation and combine computational studies with empirical studies, especially those that help to elucidate the functions of the large numbers of hypothetical and orphan genes still found in our genome databases. Likewise, the sheer vastness of these growing datasets poses myriad challenges for data management and analyses.

Table 1. A variety of resources for the study of collections of prokaryotic genomes

Excellent sources of information about a wider range of databases and web services are the special database and web server issues of *Nucleic Acids Research*, which are published every January and July, respectively (http://nar.oupjournals.org/).

PROKARYOTIC GENOMIC RESOURCES	
Monitoring completed and ongoing genome projects	
Genomes Online Database (GOLD) http://www.genomesonline.org	Provides access to lists of complete and ongoing genome projects from prokaryotes and eukaryotes
Primary international databases of complete genome sequences	
DNA Database of Japan (DDBJ) http://gib.genes.nig.ac.jp/	Genomes at DDBJ in the Genome Information Broker system
European Bioinformatics Institue (EBI) http://www.ebi.ac.uk/genomes/	Genomes at EBI
National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome	Genomes at NCBI in the Entrez Genomes system
Specialized databases	
A Systematic Annotation Package for Community Analysis of Genomes (ASAP) https://asap.ahabs.wisc.edu/annotation/php/logon.php	Genome sequences, annotations and experimental data for multiple organisms plus an interface for direct community contributions
Molligen http://cbi.labri.fr/outils/molligen/	Website dedicated to mollicute genomes allowing BLAST searching, whole-genome alignment
Oral Pathogens database http://www.oralgen.lanl.gov/	Databases of oral pathogens, bacterial and viral
Pathema http://www.tigr.org/pathema/index.shtml	In-depth curatorial analysis of pathogen genomes
STDGen and the Oral Pathogens database http://www.stdgen.lanl.gov/	Databases of genomes responsible for sexually transmitted diseases
Comparative genomic databases	
KEGG: Kyoto Encyclopedia of Genes and Genomes http://www.genome.jp/kegg/	Enzyme and pathway information about complete genomes
Comprehensive Microbial Resource (CMR) http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl	Provides access to a wide range of information and analyses about all complete bacterial genomes
Integrated Microbial Genomes (IMG) http://img.jgi.doe.gov/v1.0/main.cgi	Facilitates the visualization and exploration of genomes from a functional and evolutionary perspective
Microbial Genome Database for Comparative Analysis (MBGD) http://mbgd.genome.ad.jp/	Provides orthologue identification, paralogue clustering, motif analysis and gene order data
Virulogenome http://www.vge.ac.uk/index.html	Access to complete and incomplete genomes, including Artemis applet and ACT comparisons
Genomic feature databases	
Clusters of Orthologous Genes (COGs) http://www.ncbi.nlm.nih.gov/COG/	Individual proteins or groups of paralogues from at least three lineages corresponding to ancient conserved domains
FusionDB http://igs-server.cnrs-mrs.fr/FusionDB/	A database of bacterial and archaeal gene fusion events
Genome Atlas http://www.cbs.dtu.dk/services/GenomeAtlas/	Visualization of features within large regions of DNA; users can upload GenBank files to create custom plots
High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP) http://www.expasy.org/sprot/hamap/	HAMAP families are a collection of orthologous microbial protein families, generated manually by expert curators
Genome Reviews http://www.ebi.ac.uk/GenomeReviews/	Up-to-date, standardized and comprehensively annotated view of the genomes
Homologous Sequences in Complete Genomes Database http://pbil.univ-lyon1.fr/databases/hogenom.html	Database of homologous genes and access to phylogenetic trees
Merops http://merops.sanger.ac.uk/	Information resource for peptidases and the proteins that inhibit them

Table 1. cont.

PROKARYOTIC GENOMIC RESOURCES	
ORFanage http://www.cs.bgu.ac.il/~nomsiew/ORFans/	Access to singleton, paralogous and orthologous ORFans in bacterial genomes
OrphanMINE http://www.genomics.ceh.ac.uk/orphan_mine/	Database of bacterial proteomes with access to lists of orphans that can be filtered by a variety of criteria
Pathogenomics http://www.pathogenomics.bc.ca/IslandPathExamples.html	Identification of horizontally transferred genes and genomics islands, including pathogenicity islands
SEED http://theseed.uchicago.edu/FIG/index.cgi	Expert curation of genomic subsystems, or sets of functionally or phenotypically related genes
TransportDB http://www.membranetransport.org/	Database describing the predicted cytoplasmic membrane transport proteins
tRNAdb http://lowelab.ucsc.edu/GtRNAdb/	Genomic tRNA database which contains tRNA identifications made by the program tRNAscan-SE
Pathway and protein interaction databases	
BioCyc http://www.biocyc.org/	A collection of curated databases each of which describes the genome and metabolic pathways of a single organism
MetaCyc http://metacyc.org/	A database of nonredundant, experimentally elucidated metabolic pathways
STRING http://string.embl.de/	A database of known and predicted protein–protein interactions
Mulitple genome alignment tools	
A Genome Comparison Tool (ACT) http://www.sanger.ac.uk/Software/ACT/	A DNA sequence comparison viewer (usually BLASTN or tBLASTX) based on the Artemis genome visualization tool
Mauve http://gel.ahabs.wisc.edu/mauve/	Multiple genome alignments in the presence of large-scale evolutionary events
Multi-LAGAN http://lagan.stanford.edu/lagan_web/index.shtml	One of several packages in the LAGAN tool set for mulitple alignment of genomes
MultiPipMaker http://pipmaker.bx.psu.edu/pipmaker/	Summarizes similarity between multiple sequences using 'percent identity plots' (Pips)
Multiple Genome Aligner (MGA) http://bibiserv.techfak.uni-bielefeld.de/mga/	Computation of multiple genome alignments of large, closely related DNA sequences
Phylogenomics	
PyPhy http://www.cbs.dtu.dk/staff/thomas/pyphy/	Automatic, large-scale reconstructions of phylogenetic relationships of complete microbial genomes
Phylogenomic Display of bacterial genes (Phydbac) http://igs-server.cnrs-mrs.fr/phydbac/	Web interactive tool that displays phylogenomic profiles of bacterial protein sequences
Visualization of multiple genomes	
EnteriX http://globin.cse.psu.edu/enterix/	Visualization tools for bacterial genome alignments
Multiple Genome Navigator (MuGeN) http://www-mig.jouy.inra.fr/bdsi/MuGeN/	Tool for visual exploration of features of multiple genomes
Genomic metadata	
CMR's Genome Properties http://www.tigr.org/Genome_Properties/	Numerous attributes whose status can be described by numerical values or controlled vocabulary terms
GenomeMine http://www.genomics.ceh.ac.uk/GMINE/	Database of information about all complete genomes
Integr8 www.ebi.ac.uk/integr8/	Access to species descriptions, literature, statistical analysis and summary information about proteomes
NCBI Genome Projects http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj	Organism-specific overviews that function as portals for all projects in the database
Systematic Analysis of Completely Sequenced Organisms (SASCO) http://www.pasteur.fr/~tekaia/sacso.html	Information on base composition, amino acid composition, ancestral duplication, ancestral conservation and organisms' classification

The current status of resources for the study of collections of genomes

With the current wealth of genomes it is more compelling than ever before to find novel ways to maximize the power of comparative genomics to unravel the biology of individual isolates and groups of taxa. In the following three sections we review currently available resources for accessing complete genome sequences, databases and computational servers, and software for local data analysis.

Genome sequences. The largest collections of raw and annotated genome files can be downloaded from the genome sections of the primary international databases (Table 1). There are now also a growing number of secondary genomic resources dedicated to subsets of genomes grouped by taxonomic similarity and, more recently, by shared niche, as well as specialized projects making reannotated and added-value versions of genomes available (Table 1). All completed and ongoing genome projects can be tracked in the Genomes Online Database (Bernal *et al.*, 2001).

Databases and computational servers. Increasingly, testing hypotheses about specific genomes or sets of genes is possible using online resources that provide the ability to view and manipulate pre-computed analyses and access a range of computational tools (Table 1). Some of the genomic features that are now catalogued in online databases include: conserved orthologous genes (COGS) (Tatusov *et al.*, 2003; Uchiyama, 2003), gene fusion events (Suhre & Claverie, 2004), orphans (Siew *et al.*, 2004; Wilson *et al.*, 2005), functional groups of genes like cytoplasmic membrane transport proteins (Ren *et al.*, 2004), horizontally derived sequences (Table 1), replication origins and compositional biases (Hallin & Ussery, 2004). Among these sites, the Comprehensive Microbial Resource (CMR) (Peterson *et al.*, 2001) and the Genome Atlas (Hallin & Ussery, 2004) stand out for the variety of tools they contain.

Software for local data analysis and the creation of new tools. Although they remain a valuable resource, pre-computed datasets may not encompass a genome of interest, nor do they permit the user to explore different analytical parameters. In such cases, it is necessary to download genomes and software, and in some cases it may also be necessary to write new software and create new databases from scratch. One aid is the growing availability of toolkits that facilitate the creation of bespoke programming code, for example the widely used Bio* programming libraries (http://open-bio.org/). Another generic advance is the number of ways to quickly set up bioinformatics computing resources. These include CDs and DVDs full of bioinformatics software as well as complete turn-key workstations optimized for bioinformatics research (Tiwari & Field, 2005).

In the future it is envisioned that specialized analyses will be possible using local software that communicates with a variety of websites on the internet to process data, thus providing maximum power and flexibility. This technology is already being developed; it involves the next generation of the web, and more specifically web services and workflow tools. There are various workflow tools now becoming available, but Taverna (Oinn *et al.*, 2004) is of special interest because it is part of the MyGRID project (http://www.mygrid.org.uk/) to produce next-generation tools in support of data-intensive *in silico* experiments in biology.

Challenges and opportunities associated with the analysis of many genomes

The ability to compare more genomes is compelling from a scientific standpoint, but also brings with it a series of challenges. Issues of data storage, computational speed, file formats, integration of multiple tools, and ease of access all become more complex. On a higher level, the information required to manipulate the genomes, universal naming conventions, and the construction of databases to incorporate metadata require novel approaches. Perhaps most important are the conceptual advances that must be implemented through the development of new algorithms and statistical approaches for detecting patterns in data. For example, sequence alignment continues to be an area of active research (Miller, 2001) and approaches must now cope with the need to align millions of base pairs of sequence from two or more genomes (Table 1). The evolutionary fluidity of bacterial genomes, in terms of rapid loss and addition of genes and mobile elements and large chromosomal rearrangements, means that complete genome alignments present a far more serious computational challenge than single gene alignments writ large. Phylogenomics, for which accurate alignment is a pre-requisite, is also still actively advancing to meet the complexities involved with whole-genome data rather than single gene alignments (see http://www.phylo.org/). Here we briefly describe six research areas set to make a significant impact on our future understanding of genome sequences in a comparative genomic context.

Genomic annotation in a comparative genomic context. While the annotation of a new genome in a comparative genomic context (i.e. through comparison with an already annotated close relative) is now a common practice, and there are an increasing number of projects (re-)annotating large numbers of genomes, there are also novel ways to leverage the availability of collections of genomes to improve annotation. One practice involves successively annotating sets of functionally related genes across all genomes in a dataset instead of annotating each genome to completion before attempting another. The Fellowship for the Interpretation of Genomes consortium is developing this approach within the framework of the SEED annotation tool (Table 1), which allows experts to progressively annotate individual 'subsystems' (for example Type II secretion systems or biosynthesis of

O-glycans). This new paradigm should improve the quality and quantity of expert annotations available to the wider public.

Merging automated, experimental and curated information. As the above example illustrates, one of the most exciting prospects for the future is the comparison of information derived from automated, experimental and curated sources. This trend is underscored in the recent recommendation of the American Society for Microbiology, which recently published a report on the need to characterize functional unknown and orphan genes and build a centralized, curated database of all microbial genomes based on experimental analyses (Roberts *et al.*, 2005). Understanding the functions of the large number of hypothetical predicted proteins in our complete genome collection is one of the biggest challenges of the future and databases which help this effort will be exceptionally useful.

Visualization of genomic comparisons. The ability to summarize large volumes of genomic data in a visually intuitive format is a critical step. Currently most tools that provide access to multi-genomic information do so with respect to a single reference genome. For example, Fig. 1 was created by inputting the *Haemophilus influenzae* genome into the standalone Multiple Genome Navigator software (Hoebeke *et al.*, 2003). Databases of the future will ideally let the user switch between views based on each genome as reference strain. They will also provide novel ways to display data that expand beyond this type of view to include the ability to compare every genome with every other genome.

Dealing with multiple strains within a bacterial 'species' and leveraging the power of accessory population-level data. Within comparative genomics, special consideration should be given to the examination of multiple genomes belonging to the same named bacterial 'species' (a problematic concept in bacteria) or genus. The surprising degree of intra-species diversity revealed through genome sequencing means that a single genome sequence can no longer be viewed as defining the genetic repertoire of a named taxon, but rather as a sample of the genes potentially available to members of a given population. For example, the analysis by Welch *et al.* (2002) on three genome sequences for *Escherichia coli* revealed that only ~40 % of all identified ORFs were common to all three strains. This realization has further motivated the genome sequencing of multiple strains for a number of different named species, often in the hope that this will facilitate the genetic basis of variable phenotypes of specific ecological or clinical relevance (such as heightened virulence or antibiotic resistance).

Multiple genome sequences for single species can also reveal evidence concerning the short-term micro-evolution of bacteria, and the dynamics of genome architecture. Such analyses are most powerfully conducted within a phylogenetic or population biology framework; thus genome data should, where possible, be considered in concert with population-level data from a large number of strains (Feil, 2004). Population-level data, such as multilocus sequence typing (MLST) data (Maiden *et al.*, 1998), can reveal the degree of 'clonality' (or genotypic clustering) within populations based on sequence-level analysis of stable ('core') housekeeping genes. Alternatively evidence on the distribution of accessory elements which are likely to play a role in the adaptation to specific micro-niches can be assayed by the generation of microarray data. Microarray data can also help to identify hyper-variable regions of the genome, genomic rearrangements and evidence concerning gene expression. Furthermore, it is relatively very cheap and simple to generate data from large strain collections, although there remains an urgent need to develop more efficient software to aid in the interpretation of microarray data.

The value of taking a population genomics approach. Given the observed diversity within many named species and the increasing ease and decreasing cost of the generation of complete bacterial sequences, it seems inevitable that genomic datasets will in time encompass meaningful population samples for single species. Given an appropriate sampling regime (which could be informed by current population data such as provided by MLST), relatively small samples of sequenced strains, perhaps 10–30, will result in a powerful synthesis between genomics and population biology (Luikart *et al.*, 2003). However, in order to exploit this resource, the existing tools for the analysis of sequence alignments from single gene loci will need to be adapted to deal with complete genome sequences. The problem of alignment was dealt with earlier, but there exist a battery of statistical tests for exploring evolutionary parameters, such as the intensity or direction of selection (Yang & Bielawski, 2000), demographic changes in the population (Strimmer & Pybus, 2001) or the rate of homologous recombination (Holmes *et al.*, 1999) which could be applied to whole-genome data. A powerful approach would be to employ 'sliding windows', where data are subdivided into blocks of a user-defined size. Each block would, in turn, be analysed sequentially, thus making the computational problems of analysing large sequences tractable. Crucially, this would also provide insights into differing levels of variation along the genome and thus provide evidence as to the consistency of evolutionary forces in disparate genomic locales. For example, tests to detect homologous recombination based either on the distribution of polymorphic sites, or on the level of phylogenetic consistency, may reveal the extent (boundary points) of large-scale sequence 'mosaicism' in genomes, at the level of tens or hundreds of kilobases rather than at the level of single gene loci, the scale at which such tests are traditionally employed (Smith *et al.*, 1991). Although the possibility of large-scale homologous recombination is rarely

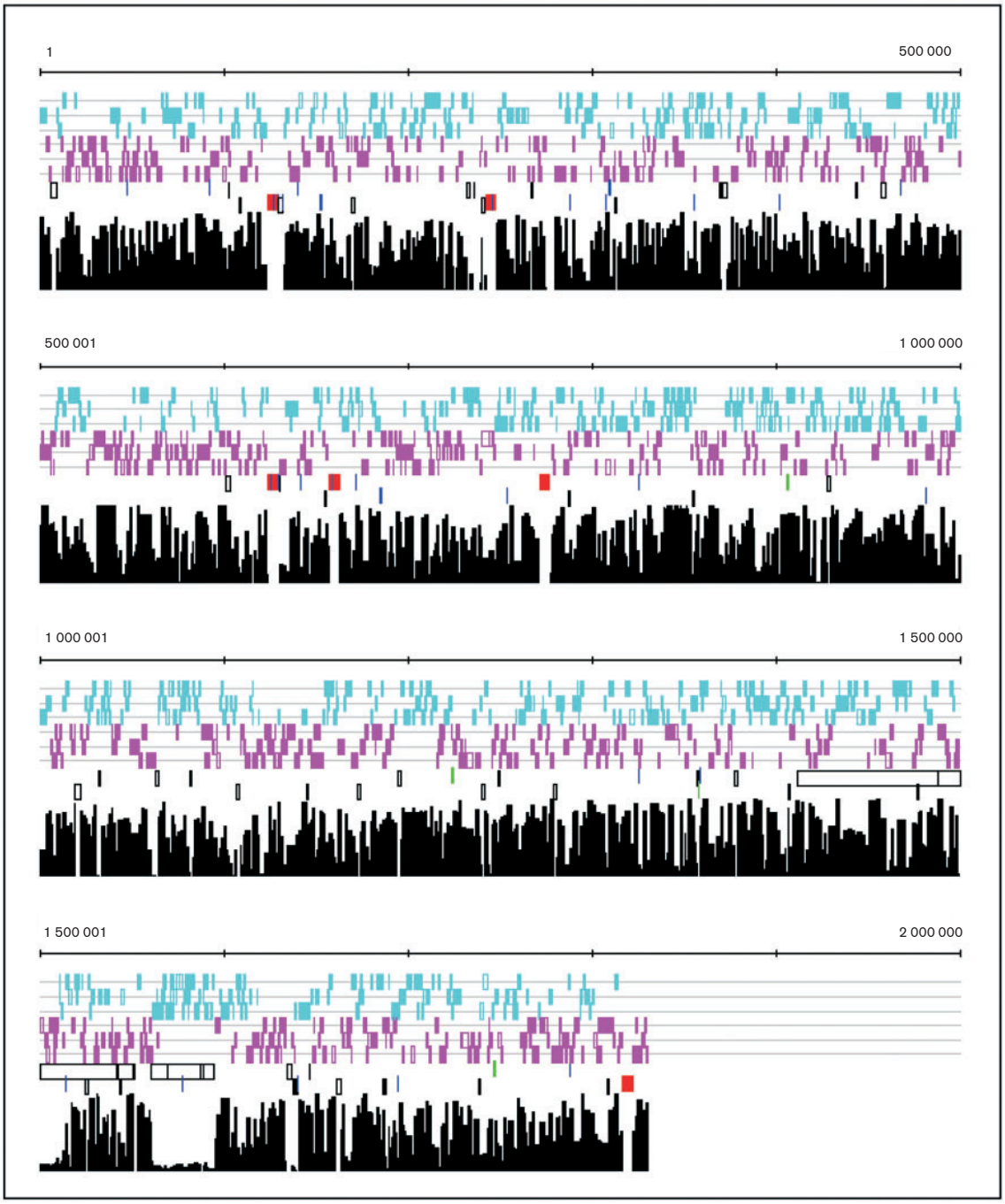

Fig. 1. A plot of the *H. influenzae* chromosome created with the Multiple Genome Navigator software. Coding regions are displayed in all six reading frames. The bottom two lines represent the forward and reverse strands and contain information such as tRNA (blue lines), rRNA (red lines), miscellaneous RNA (green lines) and miscellaneous features (empty boxes). The large empty box near the bottom represents a repeat region and the smaller box represents a phage-like area. All these features come from the GenBank file. The histogram reflects the level of conservation of each coding region. The height of the bars reflects the total number of genomes ($n=150$) in which a homologue is found.

investigated, it has been observed in *E. coli* (Guttman & Dykhuizen, 1994) and more recently in *Staphylococcus aureus* (Robinson & Enright, 2004). Thus, there is broad scope for novel types of databases and software that integrate genomes, information on mosaicism and the ability to make phylogenetic inferences (Fig. 2).

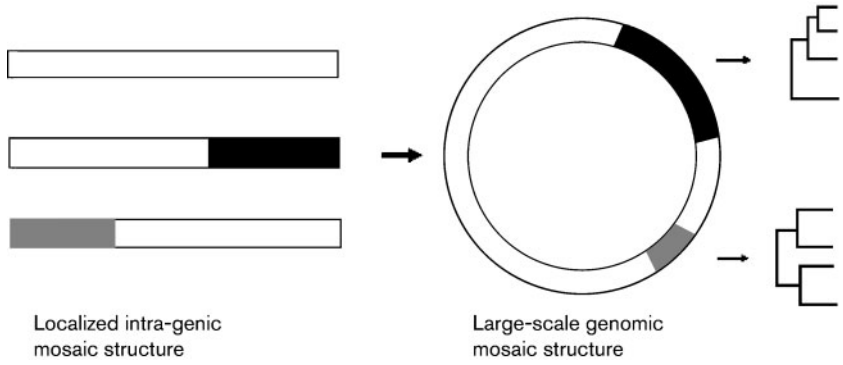

Fig. 2. Next-generation databases should merge genomic information and phylogenetic trees to unravel the mosaic nature of genomes. Localized intra-genic mosaic structure can be detected by a number of statistical procedures. These approaches could be scaled to the genome level by using a 'sliding window' approach, thus allowing the detection of homologous replacements on a much larger scale. The footprints of homologous recombination at any scale are a non-random distribution of polymorphic sites and phylogenetic inconsistency, which results in conflicting trees generated from different parts of a set of given genomes.

Mining organismal metadata to facilitate and improve large-scale comparative genomic studies. One area set to grow rapidly in coming years is that of exploring the relationships between lifestyle, evolutionary history, and genomic features. This requires access to high-quality metadata, or data that describe (or summarize) other data, in this case genomic sequences. All the information that can be used to describe a genome sequence (G+C content, number of ORFs, genome size, etc.) or the organism from which it originates (taxonomy, habitat, tropic level, etc.), is metadata and increasingly, database initiatives are adding curated organismal metadata to their content and displays. For example, the NCBI has recently launched its Genome Projects database (Table 1). Thus far, the single most impressive source of metadata for prokaryotes is the Genome Properties database within TIGR's Comprehensive Microbial Resource project (Haft *et al.*, 2005). The Genome Properties database contains over 150 fields of information for all completely sequenced bacterial genomes. While these data have been collected in part from the literature they are primarily generated through the automated analyses of homology information about particular sets of genes within genome sequences. In this way, a wealth of predicted presence or absence data on particular features (motility, pathogenicity, etc.) has been generated. A search interface is available for exploring the data and users can submit additional fields of information. Other databases use a series of tools to allow users to generate metadata that are displayed in tables and available for download. Most notably, the recently expanded functionality of the Genome Atlas also provides a wide range of analysis options and includes bacterial, viral and plasmid genomes (Hallin & Ussery, 2004).

Currently the difficulty of obtaining such metadata in a high-quality and easily accessible format is a common bottleneck in large-scale computational studies. This has led to a call for a new genomic standard (Field & Hughes, 2005) to capture information about genome sequences at the time of publication (analogous to the submission of genome annotation files). In this way, the experts generating each genome sequence would be directly responsible for providing data to the wider community about the detailed features of the organism. A catalogue of these reports would provide an extensive amount of novel data and a powerful new research tool for the future and would complement the growing number of initiatives aimed at generating computed genomic features.

Summary and a look to the future

All of the above advances rely on increased data integration and access to finer levels of detail. Ideally, next-generation databases and tools will be able to incorporate population-level data, place genomes into a rigorous phylogenetic and organismal context, and combine computed and curated data to maximize the quality and quantity of data. This will require increased interactions and collaborations between researchers working in allied fields. Multi-disciplinary approaches will be necessary to meet the challenge of interpreting the vast quantity of data generated by bacterial genome sequencing. Education will be essential if researchers are to span the cultural gulfs between these separate disciplines and work productively at the interfaces of fields like ecology and bioinformatics. A key goal will be the ability to seamlessly move between future datasets in the search for answers. Finding further ways to leverage the knowledge of our current, sizeable collection of genomes is also critical to producing rationales for the targeted study of additional taxa, genomes, populations and genes. To appreciate the challenges and potential that lie before us we need to imagine access to thousands of bacterial genomes (and tens of strains for particular species) and modify our vision of the tools we need for the future accordingly.

References

Bernal, A., Ear, U. & Kyrpides, N. (2001). Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res* **29**, 126–127.

Cole, S. T., Eiglmeier, K., Parkhill, J. & 41 other authors (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011.

Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394–1403.

Feil, E. J. (2004). Small change: keeping pace with microevolution. *Nat Rev Microbiol* **2**, 483–495.

Field, D. & Hughes, J. (2005). Cataloguing our current genome collection. *Microbiology* **151**, 1016–1019.

Field, D., Hughes, J. & Gray, T. (2005). The GenomeMine database. http://www.genomics.ceh.ac.uk/GMINE/.

Guttman, D. S. & Dykhuizen, D. E. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**, 1380–1383.

Haft, D. H., Selengut, J. D., Brinkac, L. M., Zafar, N. & White, O. (2005). Genome properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. *Bioinformatics* **21**, 293–306.

Hallin, P. F. & Ussery, D. W. (2004). CBS genome atlas database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics* **20**, 3682–3686.

Hoebeke, M., Nicolas, P. & Bessieres, P. (2003). MuGeN: simultaneous exploration of multiple genomes and computer analysis results. *Bioinformatics* **19**, 859–864.

Hohl, M., Kurtz, S. & Ohlebusch, E. (2002). Efficient multiple genome alignment. *Bioinformatics* **18 Suppl 1**, S312–S320.

Holmes, E. C., Worobey, M. & Rambaut, A. (1999). Phylogenetic evidence for recombination in dengue virus. *Mol Biol Evol* **16**, 405–409.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12.

Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet* **4**, 981–994.

Maiden, M. C., Bygraves, J. A., Feil, E. & 10 other authors (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**, 3140–3145.

Miller, W. (2001). Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* **17**, 391–397.

Oinn, T., Addis, M., Ferris, J. & 8 other authors (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**, 3045–3054.

Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K. & White, O. (2001). The comprehensive microbial resource. *Nucleic Acids Res* **29**, 123–125.

Ren, Q., Kang, K. H. & Paulsen, I. T. (2004). TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res* **32** (database issue), D284–D288.

Roberts, R. J., Karp, P., Kasif, S., Linn, S. & Buckley, M. R. (2005). *An Experimental Approach to Genome Annotation. Critical Issues Colloquia Report.* Washington, DC: American Society for Microbiology.

Robinson, D. A. & Enright, M. C. (2004). Evolution of *Staphylococcus aureus* by large chromosomal replacements. *J Bacteriol* **186**, 1060–1064.

Siew, N., Azaria, Y. & Fischer, D. (2004). The ORFanage: an ORFan database. *Nucleic Acids Res* **32** (database issue), D281–D283.

Smith, J. M., Dowson, C. G. & Spratt, B. G. (1991). Localized sex in bacteria. *Nature* **349**, 29–31.

Strimmer, K. & Pybus, O. G. (2001). Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol Biol Evol* **18**, 2298–2305.

Suhre, K. & Claverie, J. M. (2004). FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res* **32** (database issue), D273–D276.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D. & 14 other authors (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

Tiwari, B. & Field, D. (2005). A bioinformatics playground. *LinuxUser and Developer* **46**, 50–56.

Uchiyama, I. (2003). MBGD: microbial genome database for comparative analysis. *Nucleic Acids Res* **31**, 58–62.

Welch, R. A., Burland, V., Plunkett, G., 3rd & 16 other authors (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A* **99**, 17020–17024.

Wheeler, D. L., Barrett, T., Benson, D. A. & 26 other authors (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **33** (database issue), D39–D45.

Wilson, G. A., Bertrand, N., Patel, Y., Hughes, J. B., Feil, E. J. & Field, D. (2005). Orphans as taxonomically restricted and ecologically important genes. *Microbiology* **151** (in press).

Yang, Z. & Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**, 496–503.